Application of Mineral Deposit Knowledge Graph for Jilin Province Utilizing Neo4j

Xiangjin Ran[1] and Yao Pei[2,*]

[1] College of Earth Sciences, Jilin University, 6 Ximinzhu Street, Changchun, China
[2] College of Geoexploration Science and Technology, Jilin University, 6 Ximinzhu Street, Changchun, China

*peiyao@jlu.edu.cn

Abstract. A mineral deposit knowledge graph (KG) based on deep learning and natural language processing (NLP) can unveil the association between the Earth system and mineral formation. It automatically extracts the overarching model of mineral genesis and discovers mineralization laws, aiding researchers in expeditiously analyzing mineralization processes. This study constructs a mineral deposit KG for Jilin Province employing Neo4j. Initially, the Bidirectional Long Short-Term Memory (Bi-LSTM) and Conditional Random Field (CRF) model is employed to identify named entities in the text. Subsequently, the Piecewise Convolutional Neural Network (PCNN) model is employed to extract relationships between entities. Finally, the processed knowledge is imported into the Neo4j graph database for knowledge visualization. Experimental outcomes indicate that the Bi-LSTM+CRF model attains an accuracy of 91.6% and an F1 score of 90.1% in the named entity recognition task. The PCNN model reaches an accuracy of 86.4% and an F1 score of 89.3% in the relation extraction task. Through KG visualization, the correlation between controlling geological factors, prospecting indicators, and mineralization geological elements of typical mineral deposits in Jilin Province is analyzed. This provides novel technological means for the development and utilization of mineral resources in this region.

Keywords: Mineral Deposit Knowledge Graph, Deep Learning; Natural Language Processing, Neo4j Graph Database.

1 Introduction

Knowledge graphs (KGs) [1], a data organization format that has garnered significant attention in recent years, have shown promising applications across various domains. These include, but are not limited to, knowledge discovery, data analysis, the semantic web, and natural language processing (NLP) [2-3]. In the structure of KGs, entities represent abstracted real-world objects, while relationships depict the interactions between these entities. This structure encapsulates a wealth of information, including entities, concepts, attributes, and relationships. Moreover, it can be visually represent

A. E. Abomohra et al. (eds.), *Proceedings of the 2023 9th International Conference on Advances in Energy Resources and Environment Engineering (ICAESEE 2023)*, Atlantis Highlights in Engineering 29,
https://doi.org/10.2991/978-94-6463-415-0_10

ed, thereby expressing the intricate relationships of the real world in a manner comprehensible to both humans and machines. This visual representation enhances the intuitiveness of knowledge organization. Graph databases, such as Neo4j [4], are commonly employed to store and retrieve data pertaining to entities and relationships within KGs. This further underscores the practical utility and widespread adoption of KGs in contemporary data management and analysis.

Traditional entity recognition methods mainly rely on manually annotated data, which is not only time-consuming and labor-intensive, but also inefficient. Currently, there is very little research on mineral deposit knowledge graphs. Most of the research focuses on the construction and application of iron deposit knowledge graphs, and lacks exploration and analysis of knowledge graphs for other types of mineral deposits. This paper addresses this research gap by proposing a generic method for constructing mineral deposit knowledge graphs. Compared to existing methods, the innovation of this study lies in: 1) adopting the Bidirectional Long Short-Term Memory (Bi-LSTM) and Conditional Random Field (CRF) model to recognize named entities, which improves the feature extraction ability of the model; 2) The Piecewise Convolutional Neural Network (PCNN) model is used to extract entity relationships, which can effectively model semantic relationships between entities and enhance the recognition effect of cross presence relationships. This provides an effective method for constructing and visualizing knowledge graphs in the field of geology.

2 Related work

2.1 Named Entity Recognition

Labeled data forms the foundation of supervised learning models, with its quantity and quality directly influencing the predictive performance of the model. Due to the specialized nature of geological terminology, which significantly deviates from everyday language, there is currently a lack of systematic geological labeled datasets. This study adopts a named entity recognition method that combines the Bi-LSTM+CRF model [5]. The Bi-LSTM model is employed for feature extraction, while the CRF model is used for precise entity tagging. This approach enhances the recognition of named entities in ore deposit texts, compensates for the shortcomings of traditional supervised models, and results in the construction of a more robust geological knowledge extraction system.

2.2 Relationship Extraction

Entity Relationship Extraction (ERE) is a critical task within the field of information extraction. It involves the extraction of predefined relationships between entities from unstructured text, based on entity recognition. These relationships are typically represented in the form of relationship triples, denoted as <e1, r, e2>. Here, e1 and e2 represent entities, while r signifies the relationship between them. The PCNN model [6], a deep learning-based approach proposed by Zeng et al., combined with multi-instance learning methods, can be used for remote supervision of entity relationship extraction. Unlike traditional Convolutional Neural Network (CNN) models [7], the PCNN model segments sentences into three parts based on entity positions and performs pooling. This allows it to capture more context information related to entities and effectively solves the problem of error propagation in feature extraction.

2.3 Neo4j-based KG Visualization

Recent years have witnessed substantial advancements in the domain of Neo4j-based KG visualization. A myriad of node and relationship representations, layout algorithms, and interaction designs have been proposed by researchers. These contributions have laid a robust theoretical and technical groundwork for the visualization of Neo4j KGs. Neo4j Bloom [8], an open-source visualization tool specifically for Neo4j, provides a user-friendly interface that simplifies the visualization of Neo4j KGs. It encompasses numerous beneficial visualization features, such as node and relationship filtering, aggregation, and analysis. As the technology underpinning KGs continues to evolve, research on Neo4j-based visualization is anticipated to deepen further, and its application scope is expected to expand.

3 KG construction process based on Neo4j

3.1 Data Acquisition

Extracting knowledge from various sources and structures, including entities, relationships, and attributes, forms the foundation of constructing a KG. For instance, the mineral deposit model corpus established in this study primarily includes journal articles, reports, and books from CNKI. To ensure the validity of the information, it is necessary to clean the corpus, which includes steps such as word segmentation, removal of stop words, and extraction of keywords, to enhance the accuracy of the information. The collected knowledge data undergoes cleaning, normalization, extraction, and integration to enhance data quality and consistency. A unified framework is used to verify, disambiguate, and process knowledge from different sources, achieving the integration of heterogeneous data. This is crucial for the update and merge of the KG.

3.2 Ontology Acquisition

Ontology is a key input for visualizing KG models based on deep learning. This study collected text materials on the minerals of Jilin Province and constructed entity categories under the guidance of quality experts, creating a professional dataset for entity extraction in the KG. The experimental environment and parameter settings are as follows: Operating System: Windows 10; Graphics Card: Nvidia 3080Ti; Programming Language: Python 3.6; Deep Learning Platform: TensorFlow. The model parameters are as follows: Learning Rate: 0.001; Number of Hidden Units: 400; Batch Size: 64; Word Vector Dimension: 400; Dropout: 0.1; Maximum Sentence Length:

200; Number of Iterations: 100. The experimental results show that the precision of the Bi-LSTM+CRF model reached 91.6%, and the F1 score reached 90.1%, meeting the requirements of the task.

3.3 Relationship Acquisition

This step involves connecting discrete nodes by determining their relationships, thereby constructing a networked knowledge system. To ensure the accuracy of relationship extraction, this work employs the PCNN model for extracting relationships between entities. This method assumes that all possible relationship types are known, thus treating the task of relationship extraction as a text classification problem: the input is a sentence containing two entities, and the output is the relationship category between these two entities. The parameter settings for the PCNN model are as follows: Maximum Sentence Length: 200; Training Sample Size per Batch: 8; Training Epochs: 150; Learning Rate: 0.001; Dropout Rate: 0.1; Word Vector Dimension: 200; Convolution Kernel Width: 3; Number of Convolution Kernels: 200. The standard dataset for relationship extraction was divided into 70% for training, 15% for testing, and 15% for validation, which served as inputs for the neural network model. In this study, 20 types of relationships were defined. The final accuracy reached 86.4%, and the F1 score reached 89.3%, meeting the requirements of the task. These results indicate that the methods and models adopted are effective for the construction and application of KGs. The flowchart for constructing a KG is as follows (see Fig. 1):

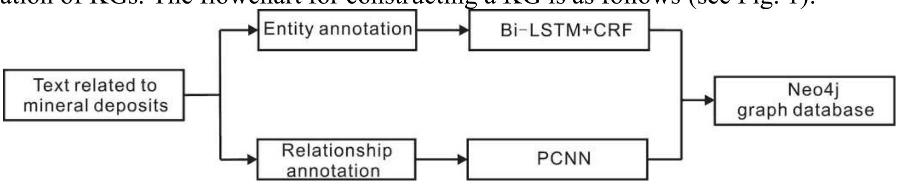

Fig. 1. Structure flowchart of KG.

3.4 Data Presentation

Data presentation is the final step in the process of constructing a KG, allowing users to intuitively view and analyze data in the KG through a graphical interface. In this study, the extracted information is transformed into structured data and visualized in the form of 'entity-relationship-entity' triples using Neo4j software. In the mineral deposit model KG, each entity corresponds to an independent conceptual ontology and exists as a node. Each entity can have zero or more attributes, which are primarily used to explain or supplement the semantic information of the node. By establishing semantic associations between nodes, 'entity-relationship-entity' triples are formed, thereby achieving knowledge visualization. The associations and constraints of knowledge have a clear directionality, where 'from' indicates the starting node of the relationship, and 'to' indicates the target node of the relationship. These relationships are represented in the form of arrows in the visualization tool (see Fig. 2).

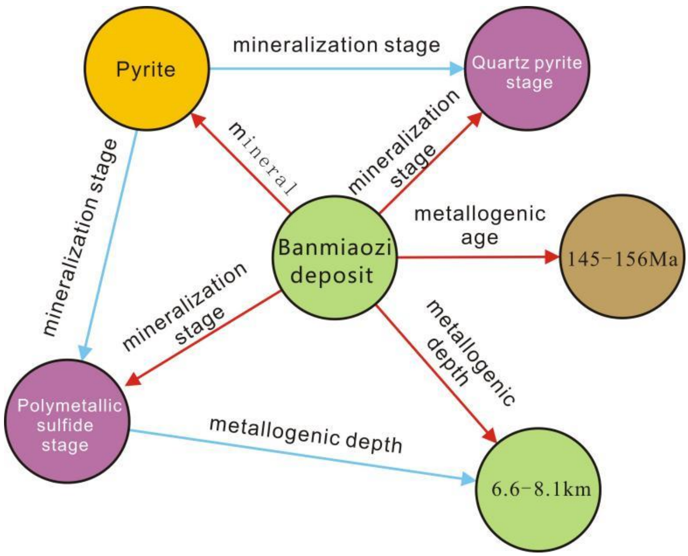

Fig. 2. Visualization of the KG.

3.5 Application of the Ore Deposit KG

This study aims to identify key indicators for mineral exploration in Jilin Province, China. The entities mainly focused on the areas of wall rock alteration, intrusive rocks, fault structures, and minerals. There was a long and complex history of magmatic activity in Jilin Province. Multiple periods and types of igneous rocks have been emplaced in the province, including ancient Proterozoic intrusive bodies and Mesozoic-Cenozoic granites and their related vein rocks. The study area is characterized by a variety of rock types, including granodiorite, porphyry, quartz monzonite, granodiorite porphyry, lamprophyre, diabase, and lamprophyre, as well as numerous gold-bearing quartz veins. In terms of spatial and temporal distribution, the relationship between vein rocks and mineral deposits is particularly important, and they are key mineralized igneous rock conditions. This indicates that the evolution of magma is closely linked to regional tectonic activity, and it also has an inseparable impact on the mineralization process.

4 Conclusions

This study successfully demonstrated the potential of mineral KGs based on deep learning and NLP technologies to reveal the relationships between the Earth system and mineral deposit formation. By automatically extracting mineral genesis models and discovering mineralization laws, the KG provides researchers with a rapid tool for analyzing metallogenic processes. Using geological features of the study area, such as stratigraphy, structure, igneous rocks, metamorphic rocks, and mineralization and alteration, combined with data from typical mineral deposits, the KG can visualize the types of mineral deposits and metallogenic geological features in the exploration area.

Acknowledgment

This work was supported by the Jilin Provincial Education Department Scientific Research Project-Knowledge and Data Collaboratively Driven Intelligent Mining of Mineral Resources Big Data in Jilin Province (Grant No. JJKH20241291KJ).

References

1. Al-Tawil M, Dimitrova V, Thakker D, Abu-Salih B.: Emerging Exploration Strategies of Knowledge Graphs. IEEE Access 11, 1 (2023).
2. Qi Z, Wang H, Zhang H.: A Dual-Store Structure for Knowledge Graphs. IEEE Transactions on Knowledge and Data Engineering. 35(2), 1104-1118 (2023).
3. Chen K, Song J, Liu S, Yu N, Feng Z, Han G, Song M.: Distribution Knowledge Embedding for Graph Pooling. IEEE Transactions on Knowledge and Data Engineering, 35(8), 7898-7908 (2023).
4. Landuyt DV, Wijshoff V, Joosen W.: A study of NoSQL query injection in Neo4j. Computers & Security 137, 103590 (2024).
5. Fucheng W, Yimin Y, Dengyun Z, Hongzhi Y, Ao Z, Guoyi C, Ma N.: Semantic Role Labeling Integrated with Multilevel Linguistic Cues and Bi-LSTM-CRF. Mathematical Problems in Engineering 2022, 1-8 (2022).
6. Xu X, Li X, Li Y, Kang L, Ge J.: A Novel Adaptively Optimized PCNN Model for Hyperspectral Image Sharpening. Remote Sensing 15(4205), 4205 (2023).
7. Nair V, Radhakrishnan A, R C, James A.: Memristive Pixel-CNN Loop Generate for CNN Generalisations. IEEE Transactions on Nanotechnology 22, 120-125 (2023).
8. Šestak M, Heričko M, Družovec TW, Turkanović M.: Applying K-vertex Cardinality Constraints on a Neo4j Graph Database. Faculty of Electrical Engineering and Computer Science 115, 459-474 (2021).